Indexing and Searching Cross Media Content in a Social Network

Pierfrancesco Bellini, Daniele Cenni, Paolo Nesi

Department of Systems and Informatics, Faculty of Engineering, University of Florence, Florence, Italy
{pbellini, cenni, nesi}@dsi.unifi.it, http://www.disit.dsi.unifi.it

Abstract—**Recent challenges in information retrieval are related to information in social networks and rich media content. In those cases, the content is associated with multilingual, user generated aspects and content, scalability, robustness and resilience to errors. Moreover, fast and efficient indexing and searching services are needed, in order to scale digital content distribution and video on demand, where huge amount of queries and content related tasks are performed by users. In this paper, an indexing and searching solution for cross media content is presented. It has been developed addressing several problems of the ECLAP social network, collecting and managing content and users in the domain of Performing Arts. The research is conducted in the scope of ECLAP (*http://www.eclap.eu*). The effectiveness of the proposed solutions has been also measured and presented.**

Keywords: content, distribution; cross media content; automated production; mpeg-21; mobile; iphone; indexing, searching, search engines, information retrieval; cultural heritage.

I. INTRODUCTION

In recent years, the rapid growth of digital resources on the Web has opened new challenges in developing efficient and robust information retrieval solutions. A wide variety of contents, with different formats and metadata types, constitute a heterogeneous set of resources difficult to deal with. An important example is given by cross media resources, which often include a rich set of metadata, mixed media, addressing serious issues when building a digital content index. Typically, there is a need of tools for metadata extraction, schemas and metadata mapping rules and tools, multilingual metadata and content translation and certification. Information retrieval (IR) systems are required to give coherent answers with respect to typos or inflexions, and must be efficient enough while sorting huge result lists. Search refinement, sorting and/or faceting techniques are major topics, especially in large multimedia repositories. Document parsing algorithms have to be fast enough in order to index high volumes of rich text documents, and to support different types of content descriptors. Indexes and repositories have to be fully accessible, without significant downtime, in case of failures or major updates of the index structure, in production services (e.g., redefinition of index schema, index corruption etc.).

Multilingual documents require query or metadata translation for information retrieval. The first approach reduces the memory usage and each document is stored only once in the index [1], while the second produces larger indexes and avoids query translation issues. Indeed the automatic query translation process could create word ambiguity, polysemy, inflection and homonymy issues [2], especially in the case of short queries [3]. Disambiguation techniques can be used, for example using co-occurrences of pair terms [4], or a general statistical approach. Query expansion [5], for example pseudo relevance feedback technique [6] [7], thesauri such as WordNet [8] or structured translation [9] can be applied to increase the retrieval efficiency. A retrieval system's effectiveness can be measured with basic estimation metrics such as precision, recall, mean average precision, R-precision, F-measure and normalized discounted cumulative gain (NDCG) [13]. Relevant test collections and evaluations series for ad hoc IR system assessment include TREC [34], GOV2 [35], NTCIR [36] and CLEF [37].

Other possible approaches for dealing with multilingual documents refer to Self Organizing Maps (SOMs) [10] or make use of sentence clustering before the translation process [11]. An alternative query translation approach involves the use of parallel or comparable Corpora [12]. They consist in a collection of natural language texts, where each document is translated in various languages; aligned parallel corpora are annotated to match each sentence in the source document with their respective translations. Thus, documents are comparable when they use the same vocabulary and deal with the same topic [13]. Ranking algorithms consist in ordering the output results list from the most to the least likely item [14]. Generally, ranking is based on location and frequency; documents with higher term occurrences are ranked higher. A notable example is the PageRank algorithm [15], which determines a page's relevance with a link analysis. Relevance feedback algorithm is based on the concept that a new query follows a modified version of the old one, derived by increasing the weight of terms in relevant items, and decreasing the weight of terms in non-relevant items. In order to overcome the limitations of traditional keyword based search engines, fuzzy approaches are used; synonyms or typos are evaluated in terms of similarity with current indexed tokens, to provide more complete results. Relevant examples of fuzzy techniques application include semantic search [16], ontologies [17], cloud computing [18], image text analysis [21], query expansion [20], clustering [19], and popular search platforms [22]. Multidimensional dynamic taxonomies models (i.e., faceted search [24] [25]) are also very popular, especially in e-commerce sites, where the user needs a way to easily explore the contents, and each facet can be represented with a taxonomy [23]. Document type detection and parsing

algorithms for metadata extraction are a valuable key factor for integrating rich text resources (e.g., semi-structured or unstructured documents) in digital indexes, with the aim of Natural Language Processing (NLP) techniques; example approaches include machine learning methods [26], table metadata extraction (e.g., from PDFs [27]), context thesauri in conjunction with document analysis [29], DOM based content extraction [28]. Typically, extracted information from unstructured documents can be organized as entities (i.e., noun phrases) and relationships between them, adjectives, tables and lists [30].

In this article, an indexing and searching solution for cross media content is presented. It has been developed addressing several problems in the area of cross media content indexing for social networks. It has been developed for the ECLAP social service portal, in the area of Performing Arts. The solution is robust with respect to typos, runtime exceptions, index schema updates, different metadata sets and content types, that constitute the ECLAP information model. It enhances and facilitates the user experience with full text multilingual search, for a large range of heterogeneous type set of content, with advanced metadata and fuzzy search, faceted access to query, content browsing and sorting techniques. The defined indexing and searching solution for ECLAP portal enabled a set of features involving a large range of rich content as: MPEG-21, web pages, forums, comments, blog posts, images, rich text documents, doc, pdf, collections, play lists, etc. Most of the activities are performed in the system back office developed with AXCP (AXMEDIS Media Gris solution for semantic computing) tools. Thus, the indexing service can be implemented in a distributed parallel architecture for massive ingestion and indexing. The proposed solution includes monitoring and logging facilities, providing data for further investigations (e.g., IR system effectiveness and user behavior assessment).

This paper is structured as follows. Section II depicts the ECLAP overview, describing the projects goals, and giving the general scenario where the solution has been integrated. It summarizes the main ECLAP tools and services. Section III presents the ECLAP indexing core services with respect to the main identified requirements and discusses the indexing related issues. Section IV depicts the ECLAP searching solution and discusses the searching strategies to increase retrieval efficiency, in the context of Performing Arts. Conclusions are drawn in Section V.

II. ECLAP OVERVIEW

The goal of the ECLAP project is to create an online archive and portal in the field of the European Performing Arts, which will also become indexed and searchable through the Europeana portal in the so called EDM data model [33]. ECLAP main objectives are to: make accessible on Europeana a large amount of Performing Arts related material as cross media content (e.g., performances, lessons, master classes, teaching material in the form of videos, audio, documents, images etc.); bring together Europe's relevant Performing Arts institutions, to provide their content on Europeana; create a stable best practice network of European Performing Arts institutions. ECLAP provides solutions and services for: Performing Arts institutions, to bridge the gap between them and Europeana, via guidelines and solutions; final users (teachers, students, actors, researchers, and Performing Arts lovers for edutainment, infotainment and entertainment). The ECLAP mission is to develop new technologies, to create a virtuous self-sustainable mechanism, to provide continuously access to the content, and to increase the number of online materials. ECLAP can be seen as a support and tool for: content aggregators (e.g., for content enrichment and aggregation, preparing content for Europeana, for content distribution); working groups on best practice reports and articles, about tools for education and training, intellectual property and business models, digital libraries and archiving. Many other thematic groups and distribution channels are also defined. ECLAP networking and social services facilities include: user group, discussion forums, mailing lists, connection with social networks, suggestions and recommendations to users, as intelligence tools (e.g., potential colleagues, using metrics based on static and dynamic user aspects, similar contents), etc. Content distribution is available toward several channels: PC/Mac, iPad and Mobiles. ECLAP includes a back office intelligence mechanisms for: automated ingestion and repurposing for metadata and content items; multilingual indexing and querying, content and metadata enrichment, IPR wizard for IPR modeling and assignment, content aggregation and annotations, e-learning support, production of suggestions. ECLAP content portal features a large set of item formats, accessible through a search service with faceting refinement and ordering. Monitoring services allow content providers to assess user data and user behavior analysis (e.g., download of contents, user satisfaction about search), reports about user preferences, with visual statistical analysis overviews on the administrative section of the portal; promotion on all indexing portals, to make more visible each partner's content. The ECLAP solution would result in cultural enrichment and promotion of European culture, in learning and research improvements. In ECLAP, users are able to deal with forums, groups, blogs, events, pages, archives, audios, blogs, braille music, collections, documents, epub, excel, flash, html, images, pdf, playlists, slides, smil, tools, videos, etc. Depending on credentials and a set of grants, each user can upload, create, improve and/or edit digital resources and their corresponding metadata. In this context, an indexing and searching service has been developed and integrated into the ECLAP Web portal.

III. INDEXING

The ECLAP content model has been designed to cope with several types of digital resources with different metadata, which require a suitable metadata mapping schema to be used for content indexing, thus enabling the whole set of content related metadata to be stored in

the same index instance. Each resource category has to map its metadata into the same set of fields, adding its specific ones into a separate set, for advanced search purposes; this results in a unified and flexible indexing schema describing the whole set of heterogeneous contents. The metadata schema is categorized in different sections (see Table I): Dublin Core (DC [32]), Technical (e.g., type of content, partner acronym providing the data, IPR model, duration, video quality, GPS position, sources, formats, etc.), Performing Arts specific metadata (such as roles of actors, relationships, information related to recording situation, etc.), ECLAP Distribution and thematic Groups, and assignment of Taxonomical terms to content. Moreover, for some of the content types, full text is accessible; comments, tags and votes may be added as user generated content.

Multilingual aspects can be at metadata level and at content body level. For example, a multilingual web page or multilingual taxonomy terms. DC and DCTerms may include a metadata language (*xml:lang* attribute that can be either mandatory, optional or not necessary). Elements that are mapped to a Performing Arts metadata are mapped to one of the generic DC/DCTerms metadata when providing metadata to Europeana. Cross media content can be MPEG-21, animations, intelligent content etc.; they share the same set of metadata, while contents that provide full text data for indexing such as blog posts, web pages, events, etc. typically do not have metadata, though they may have Groups, Taxonomical associations and comments. The ECLAP multilingual index includes the languages of the ECLAP partners (Catalan, Greek, English, Spanish, French, Hungarian, Italian, Dutch, Portuguese, Slovenian). Multilingual metadata are automatically translated from source language, and mapped into their respective language schema fields, in order to avoid query translations issues.

TABLE I. ECLAP INDEXING MODEL

MEDIA TYPES	***DC+Multilingual***	***Technical***	***Performing Arts***	***Full Text***	***Tax Group + ML***	***Comments, TAGS+ ML***	***Votes***
Cross Media (html, MPEG21, animations, etc.)	*Yn*	*Y*	*Y*	*Y*	*Yn*	*Ym*	*Yv*
Info text: blog, web pages, events, forum, comments	*T*	*N*	*N*	*N*	*N*	*Ym*	*N*
Document: PDF, DOC, ePub, ..	*Yn*	*Y*	*Y*	*Y*	*Yn*	*Ym*	*Yv*
Audio/video/image	*Yn*	*Y*	*Y*	*N*	*Yn*	*Ym*	*Yv*
Aggregations (play lists, collections, courses, etc.)	*Yn*	*Y*	*Y*	*I*	*Yn*	*Ym*	*Yv*

In Table I: *Yn* means yes with *n* possible languages (i.e., *n* metadata sets); *Yv* means yes with *v* votes; *Y* means only one set of those metadata; *I* means that the aggregation itself does not present any full text indexing, while the aggregation elements are indexed as full text entities according to their type; *T* means only title of the metadata set, *Ym* means that for that content type *m* different comments can be provided and each of them in different language. Comments may have comments as well, thus resulting in a discussion even out of a regular forum.

The above index model has been designed in order to meet the metadata requirements of the digital contents, while indexing follows the ECLAP metadata ingestion schema. 20 different partners are providing their archives by using their own custom metadata that only partially meet with standard DC. Thus, in ECLAP, digital contents are indexed in a single multilingual index, for fast access, easy management and optimization.

Catchall fields for main metadata are automatically populated at indexing time (see Table II), for full text general search (e.g., title, description, subject, content taxonomies, and Performing Arts classification), to allow multilingual metadata retrieving through a single index instance, and to build more compact Boolean queries. In this model, the catchall field includes full text of each content element. This means that in the case of Text-info media types, title and body are the only accessible fields.

TABLE II. INDEX CATCHALL FIELDS[A]

Catchall field	Component fields	Boost Weight of component fields
text	pdf_*, doc_*, docx_*, ppt_*, pptx_*, htm_*, html_*, txt_*, rtf_*	1
title	title_*	3.1
body	body_*	0.5
description	description_*	2.0
contributor	contributor_*	0.8
subject	subject_*	1.5
taxonomy	taxonomy_*	0.8
Performing Arts	PerformingArtsMetadata.#	1

A.Index fields with a _* suffix indicate language field declinations (e.g. title_english etc.)

Performing Arts metadata list

Relevant metadata are only indexed and not stored in the index, in order to keep the index smaller, easier to manage and faster to be accessed; technical metadata are stored verbatim and not analyzed (e.g., video resolution, quality, format), while descriptive metadata are fully processed (i.e., tokenized, lowercased). Metadata and structured text content from attached resources (i.e., doc, docx, ppt, pptx, xls, xlsx, pdf, html, txt) are extracted after detecting magic bytes (i.e., a prefix that identifies the file format), file extension, content type and encoding, with the aim of an internal MIME database and parsing libraries provided by Apache Tika [31]. Multilingual taxonomies are hierarchically managed in a MySQL database; each resource can be linked to a set of taxonomies, which come as a part of the resource's metadata in the index (each taxonomy path is serialized into a string, before indexing in the document structure).

Each object metadata set can be enriched and edited on the portal, to arrive at the validated/certified version to be published on Europeana. Corrections can be applied to multiple objects at the same time. Technical Metadata are used to search and retrieve them in the back office.

Avoiding downtime

The index structure is rebuilt automatically from scratch in case of corruption or major schema updates. To avoid significant service downtime, the production service keeps running while a separate instance of the index is being built, and accessory calculations are performed (e.g., taxonomy/group extraction related to each digital resource). During re-indexing, the production indexing service keeps trace of the newly uploaded contents, to index them in the new generated index; every possible thrown exception is notified by mail to the administrator, with all the relevant information (i.e., type of error, stack trace, timestamp and resource involved); mail recipients are customizable from the administrative panel settings. At the end of the re-indexing process, the newly created index is transferred over the old one.

IV. SEARCHING

The goal of the searching service is to allow the users to easily locate and sort each type of content in the ECLAP portal, and to refine their queries for a more detailed result filtering, through a fast search interface, robust with respect to mistyping; high granularity of data has to be offered to the users (i.e. advanced metadata search), with a detailed interface.

Text-Info contents (such as web pages, forums, comments, groups, events, etc.) and media contents have to be searchable in the ECLAP portal; after a query heterogeneous results may be obtained. For example, a blog post, a group, an event, a comment, a tag, a PDF, etc. In order to reduce this kind of complexity and simplify the readability of the results, in terms of play of the obtained results, the identification of comments is manifested as the presentation of the original content element, at which the comment has been provided. This means that querying for a term contained in a page, blog, forum or cross media content, has to match the set of resources containing that search term, producing a list of formatted results. Querying by taxonomy related to the content has to provide a pertinent match too. Relevance scoring has to take into account different weights for each document's metadata field; a same term occurring in different document fields is expected to provide different scoring results (i.e., a higher field's weight means a higher relevance of that field).

In order to simplify the work of users, the search service is provided as an easy to use full text, and as an advanced search. The easy to use full text frontal search is in the top center of the portal; it is offered as a text box with a search button and a drop-down menu for resource type filtering (video, audio, images, initially alphabetically ordered, while now ordered presenting on the first 5 those that are the most chosen). Queries are automatically tokenized and lowercased, before assembling the query string (i.e., a combination of weighted OR Boolean clauses, with escaping of special characters), and sent to the indexing service. Depending on the enabled languages in the ECLAP portal, each active language field is included in the query string for full text search. Advanced search is reachable from the top center portal menu, and provides language, partner and metadata filtering. The user is allowed to compose an arbitrary number of Boolean clauses in the advanced search page, thus allowing the building of a rich metadata query; for example by restricting the search to some metadata fields that only match any or all of them (OR/ALL).

In both cases, fuzzy logic facility is transparently applied to both simple and advanced search; even a query with typos can return coherent results. The query term is compared to similar terms in the index, for retrieving documents with a high degree of similarity (e.g., "*documant*" should match "*document*"), thus allowing an efficient search in case of mistyping. String metric used (Levenshtein [39] or edit distance) allows measuring the similarity between two search terms, by evaluating the minimum number of transformations needed to change one search term into another. This fuzzy similarity weight is customizable by the administrator in the portal (a weight < 1 means fuzzy logic, while a weight = 1 means Boolean logic).

In the frontal search service, a deep search checkbox is available. It allows the user to enable/disable such functionality,. thus the query string is prefixed and suffixed with a special wildcard, in a transparent way to the user, to allow searching of substrings in the index (e.g., query "*test*" matches "*testing*").

Boosting of terms is configurable on the portal. This allowed us to tune and stress the importance of certain metadata. On the basis of the performed experiments, the best appreciation has been obtained by giving more relevance to some fields with respect to others (i.e., title, subject, description, see Table II for boosting weights used in the ECLAP index). The administrator is able to change the boosting of the main search fields (i.e., title, body, description, subject, taxonomy); boost values can be extended to the whole set of metadata, though. Boosting and weighting of metadata are better tuned when the portal is more populated with significant contents. Each field of the ECLAP document structure is boosted with its predefined value at query time.

Faceted search is activated on the results of both simple frontal search and advanced search. Each faceted term is indexed un-tokenized in the ECLAP index, to accomplish a faceting count based on the whole facet. Facet parameters are appended to the query term; facet counts are evaluated from the output result by a Drupal service module, before rendering. The user can select or remove any facet in any order to refine the search. Adding or removing a facet results in adding or deleting

a search filter, and performs again the search query with or without it. Relevant facets include:

- DC: resource category, format, type, classification, creator, content language, etc.
- Technical: duration, video quality, device, publisher, source metadata language and upload time
- Group, taxonomy: genre, historical period, performing arts, coded subject

These facets can be subject to change. For instance, locations and dates, different for each historical period, can be added.

Search results are listed by relevance in descending order; this means that the first document is the most relevant with respect to the query. Results can be sorted by uploading or updating time too. The relevance is based on the occurrence of the query term in the indexed document fields: a higher number of term's occurrences (or similar terms) gives a higher score for the document. Each result item is presented with a thumbnail, relevant metadata (i.e., title and description), rating, relevance score and number of accesses; data is presented in the same language chosen by the user among the available portal localizations (or, if not available, in English or the source metadata language of the content). Results are paginated, typically 10 per page; this setting can be changed by the administrator in the settings panel. Suggestions can be enabled from the settings panel; while typing a query the system searches in the ECLAP index and may suggest similar terms to the typed one.

Search Facility Assessment

The analysis has been performed in the period from September 1st 2011 to November 30th 2011. Some of the data have been collected by using Google Analytics while others have been directly collected with internal logs. In that period, a total number of 11294 visits to the portal (of which 6032 unique visitors) have been registered. A total of 62768 views, and thus we had 5.56 pages for visit. These data were associated with 7.29 minutes of mean time of permanence on the portal. These values were substantially better than those that have been published for similar social network portals. In that period, a total of 30502 contents accesses have been registered (view, play and download, download is the 0.035%).

Moreover, Table III depicts some data about searching activities performed on ECLAP community (sorted by partnership), through queries and static menu lists available on the ECLAP portal. The numbers are referred to the same period. The first column reports the number of performed full text queries, obtaining a high rate of query per visit ratio of 37%. This means that the 37% of visitors have performed at least a query, and from these: 4051 full text queries were performed (94.56% of total), 192 faceted queries, and 41 advanced queries. Most of the queries have been performed by anonymous users. Registered users are those that are regularly registered on the portal as single users, and thus do not belong to one of the 25 institutions that have signed an agreement with ECLAP as partners or affiliated partners.

In Table III, the data related to other search results is reported to put in evidence the usage of: faceted (used only in the 5% of cases); ready-made queries to propose last posted, featured and the most popular content. The last line of the table reports the number of clicks performed on those search and list results. Thus, clicks on Last Posted Contents and Featured Contents were performed through the ECLAP menu, at the top of the home page; Clicks on Featured Contents list were performed when that list was in the home page. Moreover, it is also evident that over the 4051 queries only in 1564 cases the user has proceed to click on the provided lists of objects.

TABLE III. QUERIES / CONTENT LISTS[A]

users	# Full Text Queries	# of Faceted Queries	# Last Posted Contents	# Featured Contents	# Popular Contents
simple registered	323	24	4	22	17
Registered as partners	1094	21	27	19	9
anonymous	2634	147	234	302	213
Total	4051	192	265	343	239
Clicks after query	1564	200	318	2799	231

A. Sample Period: September 1st 2011 – November 30th 2011

It can be useful to see where the users have clicked into the lists of query results, coming from the full text and advanced query. This distribution is reported in Figure 1 in which the first 14th positions are reported. 10 of them are in the first page and the others in the first part of the second page. From these data, it can be noted that after a query on the portal, the 92.65% search results clicks were performed in the first page (first ten results). 42.27% of clicks on search results have been performed to the first proposed result. The second has received only the 14% clicks.

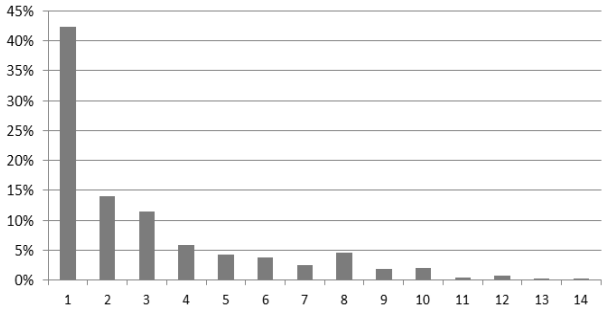

Figure 1. Clicks order distribution
(first page results and a part fo the second)

V. CONCLUSIONS AND FUTURE WORK

An integrated searching and indexing solution for the ECLAP Web portal has been developed. The ECLAP index has been designed to scale efficiently with thousands of contents/accesses; the proposed searching

facilities contribute to enhance the user experience, speed up and simplify the information retrieval process. A preliminary analysis of user sessions has been conducted, to put in evidence the user behavior on the portal. A deeper analysis is in progress to better understand the appreciation, the effective satisfaction, and user preferences.

VI. ACKNOWLEDGMENTS

The authors want to thank all the partners involved in ECLAP, and the European Commission for funding the project. ECLAP has been funded in the Theme CIP-ICT-PSP.2009.2.2, Grant Agreement No. 250481.

REFERENCES

[1] J.S. McCarley, “Should we translate the documents or the queries in cross-language information retrieval?”, in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. 1999, Association for Computational Linguistics: College Park, Maryland.

[2] Abusalah, M., J. Tait, M. Oakes, “Literature Review of Cross Language Information Retrieval”, World Academy of Science, Engineering and Technology 4, 175-177, 2005.

[3] Hull, David A., G. Grefenstette, “Querying across languages: A dictionary-based approach to multilingual information retrieval”, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR96), pages 49 – 57, Zürich, Switzerland, 1996.

[4] Yuan, S., S. Yu, “A new method for cross-language information retrieval by summing weights of graphs”, in Fourth International Conference on Fuzzy Systems and Knowledge Discovery, J. Lei, Editor. 2007, IEEE Computer Society. p. 326 - 330.

[5] L. Ballesteros, B. Croft, “Phrasal translation and query expansion techniques for cross language information retrieval”, in Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR97), pages 84 – 91, Philadelphia, PA, 1997.

[6] Attar, R., A.S. Fraenkel, “Local feedback in full-text retrieval systems”, Journal of the Association for Computing Machinery, 1977. 24: p. 397-417.

[7] L. Ballesteros, W.B. Croft, “Resolving ambiguity for cross-language retrieval”, Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval. 1998. New York: ACM Press.

[8] Fellbaum, C., “WordNet: an electronic lexical database”, Cambridge, Mass: MIT Press, 1998.

[9] Sperer, R., D.W. Oard, “Structured translation for cross-language information retrieval”, in Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR00), Athens, Greece, 2000.

[10] Lee, C.-H., H.-C. Yang, “Towards multilingual information discovery through a SOM based text mining approach”, T. Ah-Hwee and P. Yu, editors, Proceedings of the International Workshop on Text and Web Mining (PRICAI 2000), pages 80–87, Melbourne, Australia, August 28 - September 1 2000. Deakin University, Australia.

[11] Chen, H.-H, J.-J. Kuo, T.-C. Su., “Clustering and visualization in a multi- lingual multi-document summarization system”, in Proceedings of the 25th European Conference on Information Retrieval Research (ECIR 2003), number 2633 in LNCS, pages 266–280, Pisa, Italy, April 14-16 2003. Springer.

[12] Picchi, E. and C. Peters, “Cross-language information retrieval: a system for comparable corpus querying”, Cross-Language Information Retrieval, G. Grefenstette, Editor. 2000, Kluwer Academic Publishing: Massachusetts. p. 81-90.

[13] D. Manning, C., P. Raghavan, H. Schütze, “Introduction to Information Retrieval”, 2009, Cambridge: Cambridge University Press.

[14] Belkin, N. J., W. B. Croft, “Retrieval Techniques”, Williams, M. (ed.), Annual Review of Information Science and Technology, Elsevier Science Publishers, New York, 1987, Pages 109–145.

[15] Brin, S., Page, L., "The anatomy of a Large-Scale Hypertextual Web Search Engine", Computer Networks and ISDN Systems 30: 107–117, 1998.

[16] L.-F. Lai, C.-C. Wu, P.-Y. Lin, L.-T. Huang, "Developing a fuzzy search engine based on fuzzy ontology and semantic search," Fuzzy Systems (FUZZ), 2011 IEEE International Conference on , vol., no., pp.2684-2689, 27-30 June 2011.

[17] Akhlaghian, F., Arzanian, B., Moradi, P., "A Personalized Search Engine Using Ontology-Based Fuzzy Concept Networks," Data Storage and Data Engineering (DSDE), 2010 International Conference on , vol., no., pp.137-141, 9-10 Feb. 2010.

[18] Jin Li; Qian Wang; Cong Wang; Ning Cao; Kui Ren; Wenjing Lou; , "Fuzzy Keyword Search over Encrypted Data in Cloud Computing," INFOCOM, 2010 Proceedings IEEE , vol., no., pp.1-5, 14-19 March 2010.

[19] Matsumoto, T.; Hung, E. , "Fuzzy clustering and relevance ranking of web search results with differentiating cluster label generation", Fuzzy Systems (FUZZ), 2010 IEEE International Conference on , vol., no., pp.1-8, 18-23 July 2010.

[20] Takagi, T., Tajima, M., "Query expansion using conceptual fuzzy sets for search engine", The 10th IEEE International Conference on Fuzzy Systems, vol.3, pp.1303-1308, 2001.

[21] Berkovich, S., Inayatullah, M., "A fuzzy find matching tool for image text analysis", Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on , vol., pp. 101 - 105, 13-15 Oct. 2004.

[22] http://lucene.apache.org/solr/

[23] G. M. Sacco, “Research Results in Dynamic Taxonomy and Faceted Search Systems”, 18th International Workshop on Database and Expert Systems Applications, 2007.

[24] D. Tunkelang, “Faceted Search”, (Synthesis Lectures on Information Concepts, Retrieval, and Services), Morgan and Claypool Publishers, 2009.

[25] G. M. Sacco, Y. Tzitzikas, “Dynamic Taxonomies and Faceted Search”, Springer, 2009.

[26] Hui Han, C. Lee Giles, Eren Manavoglu, Hongyuan Zha, “Automatic Document Metadata Extraction using Support Vector Machines”, Proceedings of the 2003 Joint Conference on Digital Libraries, 2003.

[27] Y. Liu, P. Mitra, C. L. Giles, K. Bai, “Automatic Extraction of Table Metadata from Digital Documents”, Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, 2006.

[28] S. Gupta, G. Kaiser, D. Neistadt, P. Grimm, DOM-based Content Extraction of HTML Documents, Columbia University Computer Science Technical Reports, 2002.

[29] Shepherd, M., Watters, C. Young, J., “Context thesaurus for the extraction of metadata from medical research papers”, Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004.

[30] S. Sarawagi, “Information Extraction”, Foundations and Trends in Databases, Vol. 1, No. 3, 261–377, 2007.

[31] http://tika.apache.org/

[32] http://dublincore.org/

[33] http://www.europeana.eu/portal/

[34] http://trec.nist.gov/

[35] http://www-nlpir.nist.gov/projects/terabyte/

[36] http://research.nii.ac.jp/ntcir/data/data-en.html

[37] http://www.clef-initiative.eu/